MEMORIA
PROYECTO USISA
SEGMENTACIÓN DE CLIENTES

Andrés Membrillo Pérez

1-Introducción

Esta memoria recopila los pasos que he llevado a cabo en el proyecto realizado para la empresa Unión Salazonera Isleña S.A., nombrada a partir de ahora como USISA.

Ubicada en Isla Cristina (Huelva), USISA ha liderado el sector alimentario, consolidándose como la conservera de pescado más grande en Andalucía.

Verticalmente integrada, USISA abarca todas las actividades, desde la producción hasta el servicio postventa.

Sergio Baeza-Herrazti, representante de USISA, busca potenciar las ventas en la tienda online (www.usisa.com). Su enfoque se centra en una campaña de marketing dirigida a un grupo específico de clientes para lograr una mayor personalización y eficiencia de recursos.

2- Datos

La empresa nos proporcionó todos los datos existentes de la tienda online. Aquí nos encontramos el primer obstáculo: en 2022 la empresa cambió de sistema de base de datos. Así que existen datos registrados en dos bases de datos diferentes:

Base de datos antigua:

Formato archivo: xlsx

Periodo: Julio 2017 - marzo 2022

Shape: 10769 filas× 7 columnas

Registra por fila cada producto del carrito de compra, por ejemplo si un cliente compra 5 productos distintos, el sistema registra 5 filas con el mismo id_order.

Columnas:
'id_order', 'Cliente', 'fecha', 'Referencia del pedido ', 'product_name', 'product_quantity', 'total_price_tax_incl'.

Base de datos nueva:

Formato archivo: csv

Periodo: Marzo 2022 - **noviembre 2023**

Shape: 2019 filas × 95 columnas

Registra todos los productos del carrito de compra, en una misma fila.

Columnas:
'order_id', 'order_number', 'order_date', 'paid_date', 'status', 'shipping_total', 'shipping_tax_total', 'fee_total', 'fee_tax_total', 'tax_total', 'cart_discount', 'order_discount', 'discount_total', 'order_total', 'order_subtotal', 'order_key', 'order_currency', 'payment_method', 'payment_method_title', 'transaction_id', 'customer_ip_address', 'customer_user_agent', 'shipping_method', 'customer_id', 'customer_user', 'customer_email', 'billing_first_name', 'billing_last_name', 'billing_company', 'billing_email', 'billing_phone', 'billing_address_1', 'billing_address_2', 'billing_postcode', 'billing_city', 'billing_state', 'billing_country', 'shipping_first_name', 'shipping_last_name', 'shipping_company', 'shipping_phone', 'shipping_address_1', 'shipping_address_2', 'shipping_postcode', 'shipping_city', 'shipping_state', 'shipping_country', 'customer_note', 'wt_import_key', 'shipping_items', 'fee_items', 'tax_items', 'coupon_items', 'refund_items', 'order_notes', 'download_permissions', 'meta:_wcpdf_invoice_number', 'meta:_wcpdf_invoice_date', 'meta:_wcpdf_invoice_number_data', 'meta:_wcpdf_invoice_date_formatted', 'meta:_wcpdf_invoice_settings', 'meta:_ppcp_paypal_fees', 'line_item_1', 'line_item_2', 'line_item_3', 'line_item_4', 'line_item_5', 'line_item_6', 'line_item_7', 'line_item_8', 'line_item_9', 'line_item_10', 'line_item_11', 'line_item_12', 'line_item_13', 'line_item_14', 'line_item_15', 'line_item_16', 'line_item_17', 'line_item_18', 'line_item_19', 'line_item_20', 'line_item_21', 'line_item_22', 'line_item_23', 'line_item_24', 'line_item_25', 'line_item_26', 'line_item_27', 'line_item_28', 'line_item_29', 'line_item_30', 'line_item_31', 'line_item_32', 'line_item_33'.

3-Objetivo

El objetivo del proyecto es la **segmentación de clientes**, diferenciar y agrupar a los clientes de la tienda online según su comportamiento de compra. Con los datos existententes se pueden obtener las siguientes variables:

Frecuencia de Compra: Medición del total de pedidos realizados por un cliente en la tienda online de USISA.

Recencia de Compra: Análisis de los días transcurridos desde el último pedido efectuado por un cliente.

Gasto Total: Total(€) gastado por cliente en todos sus pedidos en la tienda online de USISA.

Los grupos tienen que estar correctamente definidos y tiene que tener sentido las variables que caracterizan a cada uno.

La hipótesis inicial es que, siguiendo el sentido común, los clientes más interesantes para la campaña de marketing son aquellos que son más habituales, realizan más pedidos y gastan más dinero. Al contrario que grupos de clientes que ya se han perdido de los cuales ya solo se puede recuperar un porcentaje

4-Limpieza y manipulación de datos

Para la segmentación de clientes hemos decidido sacrificar volumen de datos y usar exclusivamente la base de datos nueva (df_nuevo), ya que tiene muchos datos de interés sobre los clientes que son de gran utilidad, como la dirección o el correo electrónico. Los datos de la base de datos antigua los usaremos en otro proyecto.

Primero hemos limpiado el df_nuevo, que es donde se registran los pedidos por fecha:

Eliminación de columnas

Estandarizado de variables

Selección de los pedidos con status completado

Eliminación de pedidos de prueba

Eliminación de NaNs y asignación de id a cada cliente según su email. Hemos tenido que generar un nuevo id porque el sistema asignaba el id 0 a todos los clientes que han comprado sin registrarse.

Columnas df_nuevo después de limpieza:

'order_id', 'order_date', 'status', 'order_total', 'order_subtotal',
'order_currency', 'billing_first_name', 'billing_last_name',
***'billing_email'**, 'billing_address_1', 'billing_postcode',*
'billing_city', 'billing_state', 'billing_country', 'line_item_1',
'line_item_2', 'line_item_3', 'line_item_4', 'line_item_5',
'line_item_6', 'line_item_7', 'line_item_8', 'line_item_9',
'line_item_10', 'line_item_11', 'line_item_12', 'line_item_13',
'line_item_14', 'line_item_15', 'line_item_16', 'line_item_17',
'line_item_18', 'line_item_19', 'line_item_20', 'line_item_21',
'line_item_22', 'line_item_23', 'line_item_24', 'line_item_25',
'line_item_26', 'line_item_27', 'line_item_28', 'line_item_29',
'line_item_30', 'line_item_31', 'line_item_32', 'line_item_33',
***'id_cliente'**.*

Después de limpiar df_nuevo, creamos un nuevo data frame (df_cliente) donde unificamos toda la información por id_cliente, y generamos las variables que nos interesan para nuestro objetivo.

Columnas df_cliente:

'id_cliente', ***'dias_desde_la_ultima_compra'****,*
'numero_de_pedidos_por_cliente'*, 'frecuencia_compras_por_cliente',*
'facturacion_total_por_cliente(€)'*, 'order_currency',*
'billing_first_name', 'billing_last_name', 'billing_email',
'billing_address_1', ***'billing_postcode'****, 'billing_city',*
'billing_state', 'billing_country'.

5-Análisis exploratorio de los datos

Hemos realizado un análisis exploratorio de los datos de las columnas que nos interesan para la segmentación de clientes por comportamiento de compra:

- *'dias_desde_la_ultima_compra'*
- *'numero_de_pedidos_por_cliente'*
- *'facturacion_total_por_cliente(€)'*

El EDA recoge histogramas y box plot para analizar la distribución de los datos de las variables, una matriz de correlación (heatmap) para estudiar la correlación entre las variables y un mapa de pedidos localizados y agrupados por códigos postales.

Conclusiones:

Histograma y box plot: ninguna de las variables sigue una distribución normal clara, es por ello que hemos decidido normalizar los datos para el modelo.

Heatmap: las variables más correlacionadas son el número de pedidos por cliente y la facturación total por cliente(€).

Mapa: el mapa visualiza los volúmenes más grandes de pedidos en Madrid, Cataluña, Comunidad Valenciana y Andalucía.

6- Preprocesamiento de datos

Antes de aplicar el modelo normalizamos las columnas calculando su logaritmo. Y volvemos a visualizar los histogramas y box plot, para confirmar que los datos transformados siguen una distribución más normal.

Después aplicamos el escalado de los datos para transformar los datos a valores entre 0 y 1.

7- Segmentación de clientes

7.1-Clustering

Para la segmentación de clientes por comportamiento de compra, necesitamos un modelo de aprendizaje no supervisado ya que carecemos de la variable objetivo, que en nuestro caso sería el grupo o segmento que se asigna a cada cliente.

Para ello vamos a comparar dos modelos de clustering, K-Means y DBSCAN. Después de visualizar los clusters, estudiar las métricas de cada modelo, aplicamos el más eficiente.

7.2-K-Means

El modelo K.Means requiere especificar el número de clusters de antemano. Seleccionamos 3 clusters después de realizar un elbow method. Resultados para K-Means:

Silhouette Score: 0.3689
Davies-Bouldin Index: 0.9960
Calinski-Harabasz Index: 678.8144

7.3-DBSCAN

El modelo DBSCAN selecciona el número de clusters automáticamente, pero antes requiere especificar los parámetros. Seleccionamos los mejores parámetros después de hacer un grid search. El modelo identifica 3 clusters. Resultados para DBSCAN:

Silhouette Score: 0.4316
Davies-Bouldin Index: 2.2407
Calinski-Harabasz Index: 381.4763

7.4-K-Means Vs DBSCAN

Silhouette Score:

- K-Means: 0.3689
- DBSCAN: 0.4316

Un Silhouette Score más alto indica una mejor separación de los clústeres. Aunque el valor de DBSCAN es ligeramente superior, no es una gran diferencia y se debe considerar junto con otras métricas.

Davies-Bouldin Index:

- K-Means: 0.9961
- DBSCAN: 2.2408

El índice Davies-Bouldin mide la "compacidad" de los clústeres. Un valor más bajo es mejor. En este caso, KMeans tiene un valor más bajo, indicando clústeres más compactos y mejor definidos.

Calinski-Harabasz Index:

- K-Means: 678.8145
- DBSCAN: 381.4763

El índice Calinski-Harabasz mide la relación entre la dispersión dentro de los clústeres y la dispersión entre los clústeres. Un valor más alto es mejor, indicando una mejor separación de los clústeres. En este caso, K-Means tiene un valor significativamente más alto.

En general, aunque el Silhouette Score de DBSCAN es ligeramente superior, las otras dos métricas, Davies-Bouldin Index y Calinski-Harabasz Index, favorecen a **K-Means**.

Seleccionamos **K-Means** y usamos la etiqueta de los clusters con otros modelos como “RandomForestClassifier”, “LogisticRegression” y “XGBClassifier” para ver que tal predicen y explican la etiqueta arrojando una media del 98% de precisión.

7.5-Explicación de los clusters

Se realizan varios gráficos explicativos sobre los clusters, entre ellos un scatter-plot 3D que visualiza la separación de los clusters que ha realizado el modelo K-Means.

Se agrupan los clientes por sus respectivos clusters y se obtiene el centro de cada cluster usando la media para ver las características que mejor explican cada grupo.

-**Clúster 0**: su última compra ha sido hace relativamente poco, es el grupo que más pedidos promedia y que más gasta prácticamente cuadruplicando los demás. Se trata del 24% de los clientes.

-**Clúster 1**: su última compra fue hace mucho, promedian tan solo un pedido y es el grupo que menos dinero gasta. Corresponde al 54% de los clientes.

-**Clúster 2**: se observa que la última compra de este grupo ha sido muy reciente, promedian tan solo un pedido y gasta moderadamente. Comprende el 22% aproximadamente del total de clientes.

También graficamos un mapa de España con la geolocalización donde se aprecian dónde están los mayores volúmenes de pedidos de cada cluster según su código postal.

Hemos segmentado los clientes y sabemos donde se localiza cada grupo, esta información junto a los datos del cliente (correo electrónico) es de gran valor para una campaña de marketing.

8-Conclusión

El objetivo principal se logra de manera eficiente. Se distinguen los diferentes grupos de clientes además de explicar sus características. Cada cliente se identifica correctamente.

El clúster 0 lo etiquetamos como clientes fieles.
El clúster 1 lo etiquetamos como clientes perdidos.
El clúster 2 lo etiquetamos como clientes nuevos.

En cuanto a la hipótesis, se confirma que el grupo óptimo al que dedicar la campaña de marketing es el grupo de clientes fieles, que gastan mucho dinero y compran a menudo. También puede ser interesante lanzar una campaña de fidelización al grupo de clientes nuevos. En cambio no recomendamos hacer una campaña de marketing al grupo de clientes perdidos.

Frente al fenómeno de una baja retención de clientes, surge la pregunta: ¿Qué ha sucedido con el grupo de clientes que ha abandonado nuestra plataforma? ¿Existe una fuga real de clientes, o podría haber factores externos que estén influyendo en su participación activa en nuestra plataforma?